Nonresponse Adjustment in Household Surveys

Discussion paper 04007

Jelke Bethlehem & Barry Schouten

Statistics Netherlands

Voorburg/Heerlen, May 2004

Explanation of symbols

.	= data not available
*	= provisional figure
x	= publication prohibited (confidential figure)
–	= nil or less than half of unit concerned
–	= (between two figures) inclusive
0 (0,0)	= less than half of unit concerned
blank	= not applicable
2003–2004	= 2003 to 2004 inclusive
2003/2004	= average of 2003 up to and including 2004
2003/'04	= crop year, financial year, school year etc. beginning in 2003 and ending in 2004

Due to rounding, some totals may not correspond with the sum of the separate figures.

Publisher
Statistics Netherlands
Prinses Beatrixlaan 428
2273 XZ Voorburg
The Netherlands

Printed by
Statistics Netherlands - Facility Services

Cover design
WAT ontwerpers, Utrecht

Information
E-mail: infoservice@cbs.nl

Where to order
E-mail: verkoop@cbs.nl

Internet
http://www.cbs.nl

ISSN: 1572-0314
Key figure: X-10
Production code: 6008304007

Statistics Netherlands

NONRESPONSE ADJUSTMENT IN HOUSEHOLD SURVEYS

Summary: Nonresponse is a recurring problem in household surveys in many countries. Response rates of Statistics Netherlands surveys often vary between 50% and 60%. Research shows that nonresponse is usually selective. Respondents and nonrespondents differ at various demographic characteristics. To avoid a substantial negative impact on the quality of survey results, often weighting adjustment techniques are carried out. Statistics Netherlands has a large amount of background information available for this purpose. This information originates from registers and other administrative sources. The paper describes research aimed at finding auxiliary variables that are most important for including in weighting models. Also a technique is proposed to select the best weighting model. Theory is applied to data from a major survey.

Keywords: Nonresponse, Auxiliary information, Weighting

1. Introduction

1.1 About errors in surveys

It is the task of Statistics Netherlands to collect data on sources like persons, households, companies, and farms, and to transform these data into relevant, accurate and timely statistical information. Sometimes, statistical information can be retrieved from administrative sources, but more often there is a lack of such sources. In this case, a sample surveys is a powerful instrument to collect new statistical information.

In a sample survey, data are collected on only a small subset of elements in the population. Although a sample survey provides in principle only information about the sampled elements, it is possible to draw conclusions about the population as a whole. If the sample is selected using a proper sampling design, and no other problems occur, reliable estimates of population characteristics can be computed.

Estimates will never exactly equal to the population characteristics to be estimated. There will always be some error. This error can have many causes. Two broad categories can be distinguished: sampling errors and non-sampling errors.

- *Sampling errors* are introduced by the sampling design. They are due to the fact that estimates are based on a sample and not on a complete enumeration of the population. The sample is selected by means of a random selection procedure. Every new selection of a sample will result in different elements, and thus in a different value of the estimator. The magnitude of the sampling error can be controlled through the sampling design. For example, by increasing the sample

size, or by taking selection probabilities proportional to some well chosen auxiliary variable, the error in the estimate can be reduced.

- *Non-sampling errors* occur even if the whole population is investigated. Non-sampling errors are errors made during the process of recording the answers to the questions. An important source of non-sampling errors is *nonresponse*. There may be various reasons for this: refusal to co-operate, not at home at the time of the visit of the interviewer, or not able to co-operate due to illness or other circumstances.

1.2 The nonresponse problem

Nonresponse can be defined as the phenomenon that elements (persons, households, companies) in the selected sample do not provide the requested information, or that the provided information is useless. The situation in which all requested information on an element is missing is called *unit nonresponse*. If information is missing on some items only, it is called *item nonresponse*. This paper focuses on the treatment of unit nonresponse.

Due to nonresponse the sample size is smaller than expected. This leads to less accurate, but still valid, estimates of population characteristics. The confidence level of a 95%-confidence interval remains 0.95. So this is not a serious problem. It can be taken care of by taking the initial sample size larger. A far more serious problem caused by nonresponse is that estimates of population characteristics may be biased. This situation occurs if, due to nonresponse, some groups in the population are over- or under-represented, and these groups behave differently with respect to the characteristics to be investigated. If the estimate is biased, the true confidence level of a confidence interval can be much lower than assumed. For example, even if the bias is equal to the standard error of the estimate, the true confidence level of the 95%-conficence interval drops from 0.95 to 0.83, see Bethlehem & Kersten (1985). For larger sample sizes, the standard error decreases while the bias remains the same, making the situation worse.

Survey practice shows the nonresponse often causes estimates to be biased. Bethlehem and Kersten (1985) discuss a number of surveys of Statistics Netherlands. A follow-up study of the Victimization Survey showed that people, who have fear when they are alone at night, are less inclined to participate in the survey. In Housing Demand Surveys it turned out that people who refuse to co-operate, have a lesser housing demand than responding people. For the Survey on the Mobility of the Population it is obvious that mobile people are relatively under-represented among the respondents. A more recent example is an Election Survey, in which voters are over-represented among the respondents, see Voogt (2004). Indeed, estimators must be assumed to be biased unless very convincing evidence of the contrary is provided.

Bethlehem (1988) shows that the magnitude of the nonresponse bias of estimates is determined by two factors:

- The *contrast* between respondents and nonrespondents, i.e. the extent in which they differ. A large difference in the means of a survey variable for respondents and nonrespondents will lead to a large bias.
- The *response* rate. The lower the response rate is, the higher the bias will be.

Table 1.2.1 presents nonresponse figures of a number of surveys carried out by Statistics Netherlands through the years. It is difficult to compare response rates of different surveys. The magnitude of the nonresponse is determined by a large number of factors, including the subject of the survey, the target population, the time period, the length of the questionnaire, the quality of the interviewers, the fieldwork in general, etc.

Table 1.2.1. Non-response percentages of some Statistics Netherlands surveys

Year	Labour Force Survey	Consumer Sentiments Survey	Survey on Well-being of the Population	Mobility Survey	Holiday Survey
1972		29			
1973	12	23			
1974		25	28		
1975	14	22			14
1976		28	23 [2])		13
1977	12	31	30		19
1978		36		33	22
1979	19	37	35 [3])	31	26
1980		39	39	32	26
1981	17	35		32	26
1982		40	36 [2])	34	29
1983	19	37	42	34	26
1984		35 [1])		36	31
1985	23	31		39	32
1986		29	41	41	34
1987	40 [1])	29		41	
1988	41	32		45	
1989	39	32		42	
1990	39	32		45	
1991	40	31		43	
1992	43	31	55	43	
1993	43		54	44	
1994	43		48 [1])	45	
1995	40		46	46	
1996	42		48	48	
1997	44		37 [1])	50	
1998	46		40		
1999	44		40		
2000	44		43		
2001	42		40		
2003	42				
2004	41				

[1]) Complete redesign [2]) Young people only [3]) Old people only

It is clear from table 1.2.1 that nonresponse is a serious problem. Nonresponse rates in many Dutch sample surveys have increased to such a level that without special adjustment techniques one has to reckon with a decrease of the quality of the survey results. Therefore, it is vital to reduce the amount of nonresponse in the field as much as possible. Nevertheless, in spite of all these efforts, a substantial amount of

nonresponse usually remains. To avoid biased estimates, some kind of correction procedure must be carried.

1.3 The treatment of nonresponse

In practice, it is very difficult to asses the possible negative effects of nonresponse. And even if such effects can be detected, it is no simple matter to correct for them. Vital for useful detection and correction techniques is the availability of at least some information about the nonrespondents. There are two ways of obtaining more information:

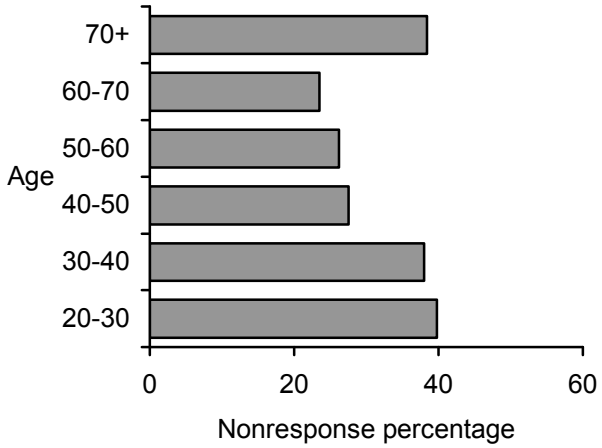
- Attempt to get information from nonrespondents. This means making contact with them, and persuading them to co-operate.
- Attempt to get information about nonrespondents. This means that no contact is required, but that information is retrieved from other sources.

A classical example of the first way is the method of Hansen and Hurvitz (1946). Already in 1946 they realized the serious implications of nonresponse. To investigate nonresponse in mail surveys, they proposed to select a random sample of nonrespondents. The selected nonrespondents were visited by interviewers, who tried to collect the required information by means of a face-to-face interview. In many cases, the interviewers were successful where the mail interview failed. Therefore, more information about possible differences between respondents and non-respondents could be obtained in this way. Moreover, this information can also be used to correct population estimates for a bias due to nonresponse. This approach of sampling nonrespondents can also be used for telephone surveys, and even for face-to-face surveys if interviewers are specially trained to persuade to nonrespondents to co-operate. A major disadvantage of the Hansen and Hurvitz approach is that it is costly and time-consuming. The fieldwork will take longer, and interviewers have to receive special training. Although, theoretically attractive, these practical problems prevent frequent application in regular surveys. A recent application of this approach can be found in Schouten and Bethlehem (2002).

Kersten and Bethlehem (1984) developed an inexpensive alternative to the Hansen and Hurvitz approach, which can be used if there is no hope of a complete interview (at that moment, or later). They call their approach the Basic Question Approach. It is based on the idea that people who refuse to answer all questions, are frequently willing to answer just a few questions. Most surveys contain a few basic questions that capture the essence of the survey. Answers of refusers to these Basic Questions provide insight in a possible bias of estimates of some important population characteristics. Moreover, the answers to the Basic Questions can be used as auxiliary information for improving the estimates for other population characteristics. The Basic Question Approach was tested in the field in the Housing Demand Survey 1981. It turned out to be rather successful, since 35% of the non-respondents were willing to answer the Basic Questions. The approach also has been successfully applied in an election survey, see Voogt (2004).

The second way of obtaining more information about nonresponse is to retrieve information from other sources. One source of information is the sampling frame. An example is the population register, which contains variables like sex, date of birth, marital status and household composition. By comparing these variables for respondents and nonrespondents, possible biases can be detected. Figure 1.3.1 shows an example based on data from a Housing Demand Survey in the city of Amsterdam. It is clear there is a relationship between nonresponse and age. Nonresponse rates are high for young people (mainly caused by not-at-home) and decrease with an increasing age. For the elderly, there is a rise in nonresponse (mainly caused by refusal).

Figure 1.3.1. Percentage nonresponse by age category



Another source of information about nonresponse is simply observation by interviewers. The kind of information that can be recorded, is type of neighbourhood, type of house, age of house, type of town, etc.

A third source of information is the statistical agency itself. By comparing the distribution of a survey variable with the available population distribution of this variable, also insight can be obtained in the effects of nonresponse. Examples of such variables are sex, age, marital status, and region. Usually, these variables are not the target variables of the survey. Still, they can be used as auxiliary variables to adjust survey estimates.

One of the most important correction techniques for nonresponse is *adjustment weighting*. It means that every responding object in the survey is assigned a weight, and estimates of population characteristics are obtained by processing weighted observations instead of the observations itself.

Weighting is based on the use of *auxiliary information*. Auxiliary information is defined as a set of variables that have been measured in the survey, and for which information on the population distribution is available. By comparing the population distribution of an auxiliary variable with its sample distribution, it can be assessed

whether or not the sample is representative for the population (with respect to this variable). If these distributions differ considerably, one must conclude that non-response has resulted in a selective sample.

The auxiliary information can also be used to compute adjustment weights. Weights are assigned to all records of responding elements. Estimates of population characteristics can now be obtained by using the weighted values instead of the unweighted values. The weights are defined in such a way that population characteristics for the auxiliary variables can be computed without error. So when weights are applied to estimate population means of auxiliary variables, the estimates must be equal to the true values.

If it is possible to make the sample representative with respect to several auxiliary variables, and if these variables have a strong relationship with the phenomena to be investigated, then the (weighted) sample will also be (approximately) representative with respect to these phenomena, and hence estimates of population characteristics will be more accurate. For an overview of weighting techniques, see e.g. Bethlehem (2002).

2. Adjustment weighting

2.1 Traditional adjustment weighting

Due to high nonresponse rates, and the possible risk of biased estimates, Statistics Netherlands has always carried out weighting on its sample survey data. In its simplest form, weighting amounts to applying post-stratification. Using qualitative auxiliary variables, the data are divided into a number of groups (strata). All cases in a group are assigned the same adjustment weight.

An early paper by Holt and Smith (1979) describes post-stratification as a robust technique that can improve the precision of estimates in case of full response. Thomsen (1973) shows how post-stratification reduces a possible bias due to nonresponse. More recently, post-stratification is discussed by Little (1986, 1993), and Gelman and Carlin (2000). Bethlehem (2002) shows that the more homogeneous the strata are, the more effective post-stratification is to reduce bias and variance.

If two auxiliary variables A en B are crossed in post-stratification weighting, this is denoted by $A \times B$. And post-stratification by crossing three variables A, B en C is denoted by $A \times B \times C$.

As more auxiliary variables are used in a weighting model, there will be more strata. Therefore the risk of empty strata or strata with too few observations will be larger. There are two solutions for this problem. One is to use less auxiliary variables, but then a lot of auxiliary information is thrown away. Another is to *collapse strata*. This means merging a stratum with too few observations with another stratum. It is important to combine strata that resemble each other as much as possible.

Collapsing strata is discussed in more detail by Kalton and Maligalig (1991) and Little (1993). Knottnerus (2003) proposes to use an estimator based on the values of inclusion probabilities of the strata.

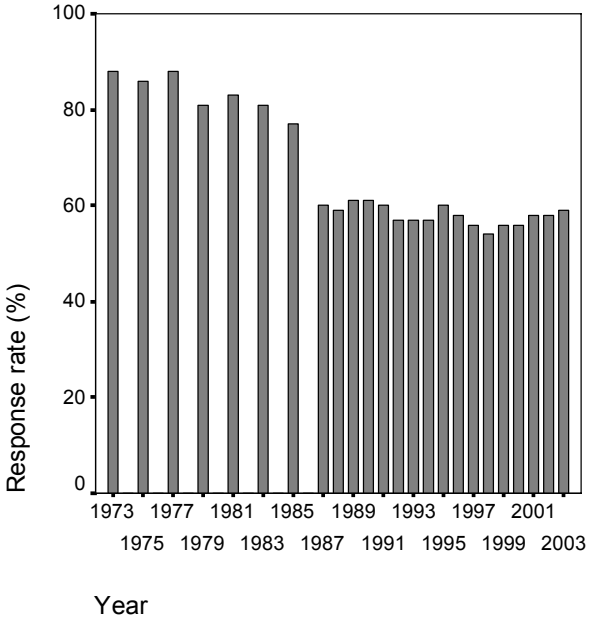
Another problem in the use of several auxiliary variables is the lack of a sufficient amount of population information. For example, if the population distributions of the two variables Age and Sex are known separately, but the distribution in the cross-classification is not known, then in this case post-stratification by Age and Sex (Age × Sex) cannot be carried out, because weights cannot be computed for the strata in the cross-classification.

To avoid these problems, and at the same time, to use as much auxiliary information as possible, Statistics Netherlands has developed and implemented other weighting techniques, like linear weighting and multiplicative weighting. See e.g. Bethlehem (1996, 2002), and Deville and Särndal (1992). For example, if post-stratification Age × Sex is not possible, it may still be possible to compute weights using the marginal distributions of both Age and Sex, but not the distribution of Age by Sex. This denoted by Age + Sex.

The number of available auxiliary variables is limited in many surveys. Very often, only demographic variables like sex, age, marital status, and region can be used. One of the most important surveys of Statistics Netherlands is the Labour Force Survey, denoted by its acronym EBB (*Enquête Beroepsbevolking*). It is an annual survey with monthly interviews. It is a 1% sample of people residing in the Netherlands, not including institutionalized people.

Figure 2.4.1 contains the response rates of the EBB over the years. In the seventies of the previous century, the response rate was almost 90%. In the eighties it dropped to a level of 60%, and never recovered.

Figure 2.4.1. Response rates in the Labour Force Survey (EBB)



The weighting model of the EBB is equal to

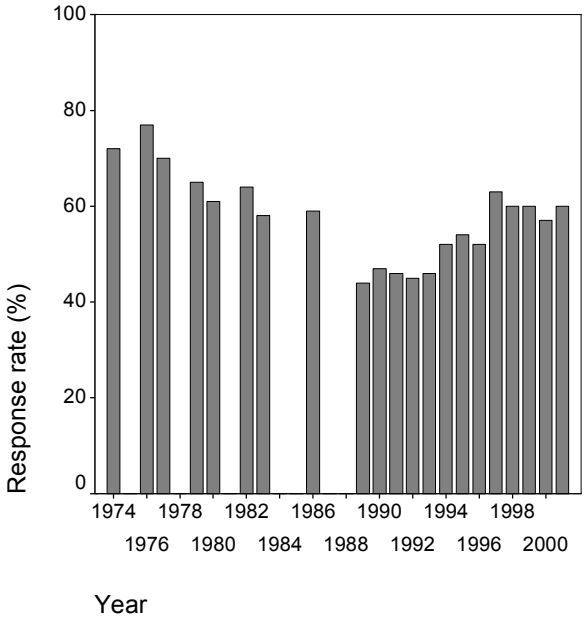
$$\text{Region}_{66} \times (\text{Age} \times \text{Sex} \times \text{Marital status})_{10} +$$
$$\text{Region}_{15} \times (\text{Age} \times \text{Sex} \times \text{Marital status})_{40} +$$
$$\text{Region}_{15} \times (\text{Age} \times \text{Sex} \times \text{Ethnicity})_{12} .$$

Subscripts denote the number of categories (strata). So weighting consist of a combination of three post-stratifications. The first one combines a detailed regional classification with a global demographic classification. The second one works the other way around: it combines a global regional classification with a detailed demographic classification. The third post-stratification resembles the second one, but marital status is replaced by ethnicity.

Another important survey is the Integrated Survey on Living Conditions. The survey will be denoted throughout this paper by its Dutch acronym POLS (*Permanent Onderzoek Leefsituatie*). POLS is a large continuous survey of Statistics Netherlands. Every month a sample is selected. The survey consists of a number of thematic modules. There is a base module with questions for all sampled persons. In addition, there are a number of modules about specific themes, such as employment situation, health, and justice. The sampled persons are distributed over these themes so that not every person has to answer questions about each theme.

Figure 2.4.2 contains the response rates of POLS over the years. Also here, a drop can be observed in the eighties of the previous century. In the early nineties, response rates are even below 50%. There was some recovery, and now response rates are around 60%.

Figure 2.4.2. Response rates in the Integrated Survey on Living Conditions (POLS)



The weighting model of POLS is equal to

$$(\text{Sex} \times \text{Age}_3 \times \text{Marital status}_2) + (\text{Sex} \times \text{Age}_{15}) + (\text{Region}_4 \times \text{Age}_3) + \text{Urbanization}_5 + \text{Region}_{16} + \text{Household size}_5 + \text{Marital Status}_4$$

The model has fewer strata than the weighting model of the EBB. This is due to a smaller sample size.

The effectiveness of a weighting procedure depends on the amount and nature of available auxiliary information. Only strong correlations between auxiliary variables and target variables, or between auxiliary variables and response behaviour help to reduce the bias. Unfortunately, correlations between these auxiliary variables and target variables in the EBB and POLS are not very strong. Consequently, the weighting models mentioned above may not be very successful in reducing nonresponse bias. Therefore, a better approach to nonresponse correction is called for.

2.2 New possibilities

To improve weighting procedures, more and more effective auxiliary variables are needed. Fortunately, some new sources have become available for Statistics Netherlands.

In the early nineties of the last century, Statistics Netherlands started the development of an integrated system of social statistics. This system is called the Social Statistics Database (SSD). The SSD will ultimately contain a wide range of characteristics on each individual in the Netherlands. There will be data on demography, geography, income, labour, education, health, and social protection. These data are obtained by combining data from registers and administrative data sources. Moreover, also data from surveys are included. These data relate to attitude, behaviour, etc. For more information about the SSD, see Everaers & Van der Laan (2001).

Using internal personal identification numbers, SSD records can be linked to survey data records, both for respondents and nonrespondents. Thus, not only demographic variables like sex, age, province of residence, and ethnicity became available for all sampled persons, but also socio-economic variables like employment and various types of social security benefits.

The Netherlands is divided in approximately 420,000 postal code areas. A postal code area contains, on average, 17 addresses. These areas are homogeneous with respect to social and economic characteristics of its inhabitants.

Using information from the population register, Statistics Netherlands has computed some demographic characteristics for these postal code areas. Since postal codes are included in the survey data file for both respondents and nonrespondents, these characteristics can be linked to the survey data file. Among the variables that can be used, are degree of urbanization, town size, percentage of people with a foreign background (non-natives). From another source also the average house value was included.

During the fieldwork period, interviewers keep record of all contact attempts. For each attempt its contact result is recorded (contact, or not). In case contact was established, the result of the co-operation request is recorded (response or nonresponse, and in case of nonresponse the reason of nonresponse). Also other information is included, like the mode of the fieldwork attempt (CAPI, CATI, etc), and whether there was contact with the person to be interviewed or another member of the household (proxy interview). All this fieldwork information can be used as auxiliary information.

Finally, two other variables turn out to be interesting as potential variables for nonresponse correction. The first one is the interviewer district code. Thus, for every respondent and nonrespondent, it is known which interviewer made the contact attempts. The second variable is an indicator whether a selected person has a listed telephone number or not.

Now that many more auxiliary variables have become available, the question becomes which of the variables may be useful in a nonresponse correction procedure, and also how they should be used in such a correction procedure. Answers to these questions have been sought in a research project in which survey data from POLS were used. Section 3 analyses the nonresponse in order to detect to what extent survey estimates may be biased. Section 4 explores a number of different weighting models that may reduce the nonresponse bias.

3. Analyses of the nonresponse in POLS

3.1 Fieldwork results.

Starting point of the nonresponse analysis was the survey data file of POLS for the year 1998. This survey data file was enriched with variables form the SSD, postal code area information, fieldwork variables, and some other variables.

Table 3.1.1. The fieldwork results of POLS in 1998

Result	Frequency	Percentage
Sample size	39 302	100,0 %
Response	24 008	61,1 %
Immediate response	9718	24,7 %
Converted refusers	14275	36,3 %
Other response	15	0,0 %
Nonresponse	15294	38,9 %
Unprocessed cases	2514	6,4 %
Not able (illness, handicap)	735	1,9 %
Persistent refusers	8918	22,7 %
No contact	2093	5,3 %
Language problems	416	1,1 %
Moved	376	1,0 %
Other nonresponse	242	0,6 %

The fieldwork of POLS covered a period of two months. In the first month, selected persons were approached with CAPI (*Computer Assisted Personal Interviewing*). For persons that could not be contacted or refused and who had a listed phone number, a second attempt was made in the second month using CATI (*Computer Assisted Telephone Interviewing*). For persons without a phone or without a listed phone number, a second attempt was made with CAPI.

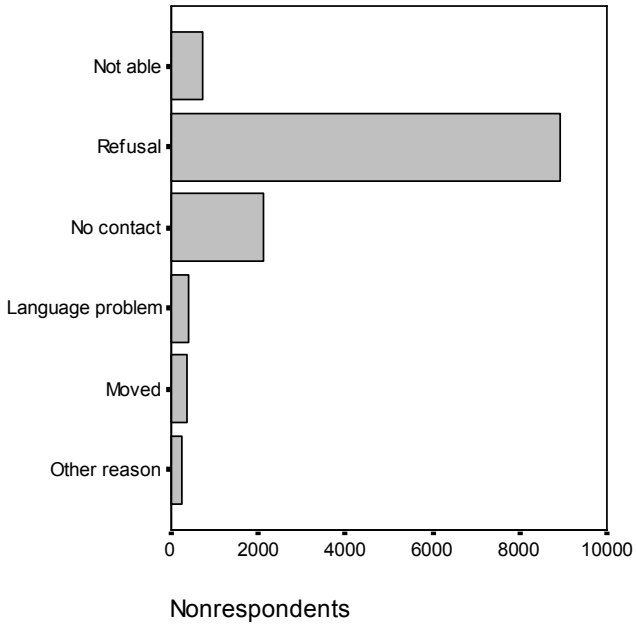
There were 2,514 cases selected in the sample that were not processed by the interviewers in the field. Leaving out these cases results in a net sample of 39,302 – 2,514 = 36,788 cases. In 24008 cases there was a response. So, the response percentage was 65.3%

The POLS survey distinguishes the following main groups of nonrespondents:

- Persons not able to participate due to illness or handicap
- Refusers
- Persons that could not be contacted
- Persons with language problems
- Moved persons

Figure 3.1.1 shows how big the various nonresponse groups are. Clearly, the most important group was that of the refusers (69.8%). The non-contact group was much smaller (16.4%). The third group (5.8%) consisted of the not-able.

Figure 3.1.1. Composition of the nonresponse



Language problems (3.3%) and moved persons (2.9%) constituted only a very small portion of the total nonresponse.

The unprocessed cases and various nonresponse groups are analysed in some more detail in the following sub-sections. A more detailed account can be found in Bethlehem en Schouten (2004).

3.2 Unprocessed cases

There were 2,514 cases selected in the sample that were not processed by the interviewers in the field. Reasons for this type of nonresponse are lack of capacity (too high workload for the interviewer) and interviewer not available (illness, holidays). Since there is a substantial amount of this so-called *administrative nonresponse*, it is important to find out whether this may be the cause of any bias.

The unprocessed and processed cases were compared for the available auxiliary variables. For the following explanatory variables significance differences were found between processed and unprocessed cases:

- Marital status of selected person;
- Percentage of people in the area of residence with a foreign background (non-natives);
- Province of residence;
- Size of town of residence;
- Degree of urbanization of town of residence;
- Whether or not the selected person has a listed telephone number;
- Interviewer district.

The logistic regressions showed that the highest explanatory power (9.1%) could be obtained by a model having interviewer district and average house value. Replacing house value by percentage of non-natives had the same effect. Adding more variables did not show significant improvement. The results of the logistic regressions are summarized in table 3.2.1.

Table 3.2.1. Results of logistic regressions for processed/unprocessed

Dependent variable	Explanatory variables	Percentage of explained variance
Processed/unprocessed	Province	6.4 %
	Interviewer district	8.9 %
	Interviewer district, Average house value	9.1 %
	Interviewer district, Percentage non-natives	9.1 %

To obtain more insight in the possible relationship between whether or not a case has been processed and auxiliary variables, also a classification tree was constructed. Such a technique breaks up the data set into a number of homogeneous groups by

means of a succession of binary splits. Thus, groups of cases are obtained that are increasingly more homogeneous with respect to the response variable. A more elaborate application of this technique can be found in Mesa et al. (2000).

The conclusions from classification tree analysis were similar to those of the logistic regression, but more detailed. For example, one could see that the highest percentage of unprocessed cases (28.8%) is found for cheap houses in the province of North-Holland and the town of Amsterdam. And the lowest percentage of unprocessed cases (1.5%) can be found in the rural areas of the province of North-Holland.

3.3 Moved persons

After the sample had been selected, Statistics Netherlands asked the municipalities to provide recent address information of the selected persons. Municipal registrations are not completely up-to-date. They lag behind a few months with respect to moving. Therefore, it may happen that a selected person has moved to another address. In theory, if interviewers are able to determine the new address, e.g. by asking the new inhabitants, and this new address is within their region, they have to visit the new address. In practice, this almost never happens.

From table 2.1.1 it is clear that the moved persons form a relative small group of nonrespondents. They constitute about 1% of the total sample size.

The logistic regressions showed that highest explanatory power (15.3% explained variance), was obtained in a model having age, marital status, and degree of urbanization as explanatory variables, see table 3.3.1.

Table 3.3.1. Results of logistic regressions for moved/not moved

Dependent variable	Explanatory variables	Percentage of explained variance
Moved / not moved	Age	11.5 %
	Age, Marital status	14.3 %
	Age, Marital status, Degree of urbanization	15.3 %

The conclusion is that traditional auxiliary variables may be used to correct for a possible nonresponse bias.

A classification tree analysis showed that another interesting variable is the composition of the household. The rate of moving is lowest (0.2%) among married couples, single parents and unmarried couples without children who are older than 33 years. The rate of moving of highest (6.6%) among single persons and unmarried couples of an age of between 17 and 37 years with children and who live in certain neighbourhoods with a high percentage of non-natives.

3.4 Hard-to-contact persons

The fieldwork of POLS 1998 was spread over two months. In the first month, contact attempts were made with 34,084 persons. Contact was established in approximately 90% of the cases.

In case of nonresponse in the first month, new contact attempts were made by telephone in the second month. This nonresponse group consisted of 7,895 persons. Contact was established in 84% of the cases.

In total, for the whole fieldwork period, there have been contact attempts with 36,412 persons. These attempts were successful in 94% of the cases. For 2,093 persons no contact at all could be established. This constitutes 5.3% of the sample.

The contacted and not contacted persons were compared for the available auxiliary variables. The logistic regressions showed that highest explained variance (15.2%) was obtained in a model containing listed phone number (yes/no), interviewer district, age by marital status, average house value, and household size, see 3.4.1.

Table 3.4.1. Results of logistic regressions for contact/no contact

Dependent variable	Explanatory variables	Percentage of explained variance
Contact / no contact	Listed phone	6.7 %
	Listed phone, Interviewer district	11.0 %
	Listed phone, Interviewer district, Age by marital status	13.8 %
	Listed phone, Interviewer district, Age by marital status, Average house value	14.2 %
	Listed phone, Interviewer district, Age by marital status, Average house value, Household size	15.2 %

An interesting question is whether no-contact cases resemble hard-to-contact cases. If so, the hard to contact persons can provide information about the no contacts. To answer this question, it is important to define what a hard-to-contact person is. One way to do it is to define it in terms of contact probabilities. The contact probability is the probability of success in a contact attempt at a randomly selected moment during the fieldwork period.

For each selected person, the contact probability could be estimated if a number of independent random contact attempts could be made. This can be seen as a series of Bernoulli trials with probability of success being equal to the contact probability.

Unfortunately, the fieldwork of POLS 98 was not organized in such a way that it allowed for easy estimation of contact probabilities. The contact attempts were

certainly not chosen at random moments. Also, the number of contact attempts changed from address to address.

A classification tree analysis for hard-to-contact persons showed that the percentage of hard-to-contact persons was lowest (2.9%) for the group with a listed phone number, and consisting of children, married people, single parents, and partners in unmarried couple without children. The percentage of hard-to-contact persons was highest (30.3%) for people without a listed phone number in certain interviewer districts in the four big cities.

3.5 Language problems

In about 1.1 % of the sample cases response was not possible due to language problems.

The Dutch speaking and non-Dutch speaking persons were compared for the available auxiliary variables. It did not come as a surprise that two auxiliary variables turned out to be correlated with the language problem variable. The first of these variables was ethnic origin (Native Dutch, Moroccan, Turkish, Surinamese, Antillean, Other western origin, Other non-western origin). The second variable was ethnic generation (Native Dutch, First generation non-native, Child of non-native parents, Child with one non-native parent).

The logistic regression model with age and a crossing of ethnic group and ethnic origin had the highest percentage of explained variance: 51.2%. The results of the logistic regressions are summarized in table 3.5.1.

Table 3.5.1. Results of logistic regressions for Dutch/non-Dutch speaking

Dependent variable	Explanatory variables	Percentage of explained variance
Dutch speaking / not Dutch speaking	Age, Ethnic generation	45.1 %
	Age, Ethnic origin	46.5 %
	Age, Ethnic generation, Ethnic origin	51.2 %

There is some indication that employment status also has some explanatory power. However, the number of observations was too small for a definite conclusion.

Language problems are most serious for first generation Moroccans and Turks. The percentage of nonresponse due to language problems is 29% for the Turks, and 25% for Moroccans.

3.6 Refusals

After subtracting the no-contacts, the group of people remains for which contact was established. These people were invited to participate in the survey. The response rate

in this group was 71.3%. A large part of the response (79.6% of the 71.3%) was already realized in the first month.

It should be noticed that the contacted group is not representative for the target population, because the no-contacts have been removed from it.

First, the refusals in the first month were analysed. The response rate among the contacts was 62.4%.

The participants and refusers were compared for the available auxiliary variables. Nearly all auxiliary variables had a significant contribution in these models. Due to the large sample size, even small contributions turned out to be significant. There were only very small differences in participation rates for the various categories of the auxiliary variables, with a number of exceptions:

- Participation rates for people under 18 years were much higher than rates for older people;
- There were geographic differences. Co-operation rates in the big cities are much lower.
- Participation rates were much higher for persons with a listed telephone number. People with a non-listed telephone number tended to refuse more. Note that in the first month the listed telephone numbers did not play a role in the fieldwork strategy. So, this is not a fieldwork effect.
- Participation rates among Moroccans and Turks were higher. The on average lower response rate for these groups are apparently caused by language problems, and not by refusal.

Separate analyses were carried out for the group of age 18 year and older in the big cities, and for the group of age 18 and older in other parts of the country.

In the big cities, the variables with most explanatory power were the percentage of non-natives in the neighbourhood and the average house value. Participation decreases with an increasing percentage of non-natives and a decreasing housing value. So there seems to be a relationship between participation and social-economic status of the neighbourhood.

Outside the big cities, the interviewer district turned out to have most explanatory power. Apparently, the performance of interviewer plays a significant role here.

A second analysis was carried out on refusals in the complete fieldwork period. Persons below the age of 12 were not included in this analysis. They have different inclusion probabilities. Consequently, percentages in the results reflect true response rates.

A classification tree analysis showed that refusal rates in groups vary between 19.9% and 58.2%. Refusal rates under 20% are found for two groups:

- The rate was 19.3% for people with an age of at least 50 in households consisting of at least four members, and with a listed telephone number;

- The rate was 19.9% for people in certain districts, with an age of at most 18 years, and who have a listed phone number.

Refusal rates of over 40% were found for three groups:

- The rate was 58.2% for people without a listed phone number in the town of The Hague who live in small households (less than four members);
- The rate was 42.1% for people without a listed phone number in other parts of country, and who live in small households (less than four members);
- The rate was 40.0% for people without a listed phone number in larger households (more than three members) in certain districts.

Particularly, there were strong effects for age and listed telephone number. Among the people of age 18 years and older, the variable listed telephone number (yes/no) had the highest explanatory power. The effect was even stronger than in the first month. This does not come as a surprise because of the role telephones play in the second month of the fieldwork.

It should be mentioned that response rates for first generation non-natives decrease in the second month of the fieldwork. This is due to the fact that telephone ownership among this group is much lower than for the rest of the population.

3.7 Not able to participate due to illness or handicap

There is a small group of 735 (1.9%) people for which no response was obtained due to illness or handicap. In 416 cases there had been no direct contact between the interviewer and the selected person. Apparently, contact was established with another person in the household. From the fieldwork reports it was not clear whether illness or handicap was the direct cause of nonresponse. It is not unlikely that illness or handicap was used as a reason for nonresponse to avoid having to refuse.

Comparison of the not able persons with other persons showed a strong relationship with age, and a somewhat less strong relationship with marital status, having a job, receiving a social benefit due to disablement.

3.8 Conclusions from the analysis

The analysis of the POLS 1998 data showed that additional auxiliary variables help to explain what is going on with respect to response and nonresponse. Not only demographic and socio-economic variables are useful in this respect, but also fieldwork variables that describe various contact attempts.

Traditionally, fieldwork reports are made to monitor fieldwork and interviewer performance. Use of this type of information in a nonresponse analysis requires this information to be recorded in a more systematic way. Also, it is important that fieldwork information becomes a standard part of the survey data file.

The analysis of the nonresponse of POLS 1998 showed that it also helps to have the following auxiliary information available for both respondents and nonrespondents:

- Whether or not a person has a listed phone number.
- The composition of the household, and other characteristics of the household.
- Characteristics of the house
- Socio-economic status of the neighbourhood

4. Selection of weighting models

4.1 Introduction

One problem that many adjustment methods have in common is the selection of auxiliary variables. Nonresponse theory suggests using auxiliary variables to form homogeneous groups, see e.g. Bethlehem (2002). If these groups are homogeneous with respect to the target variable, bias will be reduced. The same happens if groups are homogeneous with respect to response behaviour (response probabilities). In practice, this approach is implemented in two steps:

1) Auxiliary variables are selected that are correlated with response behaviour. This means that response rates are different in different strata;
2) If many auxiliary variables are selected, the weighting model may become too big. Then, there will be a risk of strata with too few observation, resulting in unstable estimates. In this situation the set of auxiliary variables may be reduced by removing variables that are uncorrelated with the target variables.

Other approaches are possible. Little (1986) proposes to form so-called adjustment cells by modelling the response probability, forming response groups and clustering response groups based on the differences between the 'average' answers to the survey questions. See also Rosenbaum and Rubin (1984), Ekholm and Laaksonen (1991) and Czajka et al. (1992). Eltinge and Yansaneh (1997) compare several criteria for the formation of adjustment cells.

Geuzinge, Van Rooijen and Bakker (2000) propose to use the product of the correlation between the response indicator and the auxiliary variables, and the correlation between a target variable and the auxiliary variables as a measure for the relevance of auxiliary variables in a weighting model.

Crucial for an effective nonresponse adjustment are the assumptions made about the nonresponse mechanism. The nonresponse mechanism is called Missing Completely at Random (MCAR) whenever the probability of response is independent of the survey questions. In case the probability of response is independent of the survey questions when conditioned on a set of auxiliary variables, the mechanism is called Missing at Random (MAR). For most surveys the MCAR assumption does not hold for the auxiliary variables. In practice it is usually assumed that the nonresponse can be made 'sufficiently' MAR by incorporating available auxiliary variables in a weighting model.

Schouten (2003) presents a method to compute a lower and upper bound for the maximal absolute bias of the regression estimator. He uses this result to propose a selection technique for auxiliary variables that minimizes the width of the interval for the bias. This selection strategy is explained in section 4.3. First, the general regression estimator is introduced in section 4.2.

4.2 The generalized regression estimator

Let the finite *survey population* U consist of a set of N identifiable elements. The values of target variable Y for these elements are denoted by Y_1, Y_2, …, Y_N. Objective of the sample survey is assumed to be estimation of the population mean

$$\overline{Y} = \frac{1}{N}\sum_{k=1}^{N} Y_k \tag{4.2.1}$$

To this end, a sample of size n is selected. It is assumed throughout this chapter that samples are simple random without replacement. Let y_1, y_2, …,y_n denote the values of the target variable for the n selected elements.

In case of complete response, the sample mean

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_k . \tag{4.2.2}$$

is an unbiased estimator of the population mean.

To analyse the impact of nonresponse on the characteristics of estimators, the phenomenon of nonresponse must be incorporated in the theory of sampling. Here, the *Random Response Model* is assumed. This model assigns to each element k in de population a certain, unknown probability ρ_k of response when contacted in the sample.

Let

$$\overline{y}^{*} = \frac{1}{m}\sum_{i=1}^{m} y_i \tag{4.2.3}$$

denote the mean of the m (m < n) available observations. The obvious approach would be to use this estimator to estimate the population mean. Unfortunately, this is not an unbiased estimator. Bethlehem (1988) shows that the expected value of this estimator is approximately equal to

$$E(\overline{y}^{*}) \approx \frac{1}{N}\sum_{k=1}^{N} \frac{\rho_k Y_k}{\overline{\rho}} . \tag{4.2.4}$$

Therefore, the bias of this estimator is approximately equal to

$$B(\overline{y}^{*}) \equiv E(\overline{y}^{*}) - \overline{Y} \approx \overline{Y}^{*} - \overline{Y} = \frac{C_{\rho,Y}}{\overline{\rho}} = \frac{R_{\rho,Y} S_\rho S_Y}{\overline{\rho}} , \tag{4.2.5}$$

in which $C_{\rho,Y}$ is the population covariance between response probabilities and values of the target variable, $R_{\rho,Y}$ is the population correlation coefficient between response probabilities and values of the target variable, S_ρ is the standard deviation of the response probabilities, and S_Y is the standard deviation of the target variable.

Hence, the estimator is unbiased if there is no correlation between the target variable and response behaviour. The stronger the relationship between the target variable and response behaviour, the larger the bias. The size of the bias also depends on the amount of nonresponse. The more people are inclined to co-operate in a survey, the higher the average response probability will be, resulting in a smaller bias.

Schouten (2003) shows that, for any auxiliary variable Z, the correlation coefficient $R_{\rho,Y}$ always lies in the interval

$$R_{\rho,Z}R_{Z,Y} \pm \sqrt{1-R^2_{\rho,Z}}\sqrt{1-R^2_{Z,Y}} \tag{4.2.6}$$

Therefore, the bias of estimator (4.2.3) always lies in the interval

$$\frac{S_\rho S_Y}{\bar{\rho}}\left(R_{\rho,Z}R_{Z,Y} \pm \sqrt{1-R^2_{\rho,Z}}\sqrt{1-R^2_{Z,Y}}\right), \tag{4.2.7}$$

for any auxiliary variable Z.

In the case of full response, the precision of the Horvitz-Thompson estimator can be improved if suitable auxiliary information is available. Suppose, there are p auxiliary variables. The values of these variables are for element k denoted by the vector $X_{k1},X_{k2},\ldots,X_{kp}$. We assume that $X_{k1}=1$ for all k, i.e. there is a constant term in the regression model.

The vector of population means is denoted by $\overline{X}$. If the auxiliary variables are correlated with the target variable, it is possible to find a vector $B = (B_1,B_2,...,B_p)'$ of regression coefficients for a best fit of Y on X, for which the residuals $E_k = Y_k - X_kB$ vary less then the values of target variable itself. Application of ordinary least squares theory gives

$$B = \left(\sum_{k=1}^{N} X_k X_k'\right)^{-1}\left(\sum_{k=1}^{N} X_k Y_k\right). \tag{4.2.8}$$

In the case of full response, this vector of coefficients can be estimated by

$$b = \left(\sum_{i=1}^{n} x_i x_i'\right)^{-1}\left(\sum_{i=1}^{n} x_i y_i\right), \tag{4.2.9}$$

where y_i and x_i denote the sample values. The estimator b is an asymptotically design unbiased (ADU) estimator of B. This means that b is approximately unbiased for large samples. Using (4.2.9) the *generalized regression estimator* for the case of full response is defined by

$$\bar{y}_{REG} = \bar{y} + (\bar{X} - \bar{x})'b, \tag{4.2.10}$$

in which $\bar{x}$ is the vector sample means of the auxiliary variables. The generalized regression estimator is an ADU estimator of the population mean of the target variable. The estimator and its properties are discussed by a number of authors, see e.g. Bethlehem (1988), Isaki and Fuller (1982), and Robinson and Särndal (1983).

It can be shown that *post-stratification* is a special case of the generalized regression estimator. To that end a vector X of L dummy variables is introduced, where L is the number of strata. The first dummy always assumes the value 1 (corresponding to the constant term in the regression model). The remaining L-1 dummies correspond to the first L-1 strata. The value X_{kh} of h-th dummy variable for element k assumes the value 1 if element k is in the stratum, and otherwise it assumes the value 0 (for h=1, 2, …, L-1). In this situation, the generalized regression estimator reduces to the post-stratification estimator.

Until now, the general regression estimator was described for the situation in which no nonresponse occurred. This estimator can also be applied in the case data are missing completely at random (MCAR). For this type of nonresponse, the available data are a random sub-sample of the data for the sampled elements. In case data are not missing completely at random, the estimator will be biased. Therefore, a modified regression estimator is defined based on modified Horvitz-Thompson estimators:

$$\bar{y}^*_{REG} = \bar{y}^* + (\bar{X} - \bar{x}^*)'b^*, \tag{4.2.11}$$

in which $\bar{y}^*$ is defined by (4.2.3), $\bar{x}^*$ is the analogue of $\bar{y}^*$, and b^* is defined by

$$b^* = \left(\sum_{i=1}^{m} x_i x_i' \right)^{-1} \left(\sum_{i=1}^{m} x_i y_i \right) \tag{4.2.12}$$

Bethlehem (1988) shows that, if there exists a vector c of fixed numbers such that X_k'c = 1 for all X_k, the bias of the modified regression estimator is approximately equal to

$$B(\bar{y}^*_{REG}) = \bar{X}B^* - \bar{Y}, \tag{4.2.13}$$

where B^* is defined by

$$B^* = \left(\sum_{k=1}^{N} \rho_k X_k X_k' \right)^{-1} \left(\sum_{k=1}^{N} \rho_k X_k Y_k \right). \tag{4.2.14}$$

From (4.2.13) it is clear that the bias vanishes if $B^* = B$. Thus, the regression estimator will be unbiased if nonresponse does not affect the regression coefficients. By writing

$$B^* = B + \left(\frac{1}{N} \sum_{k=1}^{N} \frac{\rho_k X_k X_k'}{\bar{\rho}} \right) \bar{E}^*, \tag{4.2.15}$$

where

$$\overline{E}^{*} = \frac{1}{N}\sum_{k=1}^{N}\frac{\rho_k E_k}{\overline{\rho}}, \tag{4.2.16}$$

two conclusions can be drawn. First, B^{*} and B will be approximately equal if quantity (4.2.16) is small. So a good fit of the regression model will result in small residuals, and thus will reduce the bias. Second, B^{*} and B will be approximately equal if (4.2.16) is close to or equal to 0, and this will be the case if there is little or no correlation between the residuals of the regression model and the response probabilities. It can be shown that in case the data are missing at random (MAR), and the auxiliary variables concerned are included in the regression model, quantity (4.2.16) will indeed be equal to 0.

4.3 Selection of auxiliary variables

The previous section made clear that the general regression estimator can be used as a framework to reduce nonresponse bias. To select the most effective auxiliary variables in the weighting model, a strategy is proposed that minimizes the maximal absolute bias under a weaker assumption than MAR. This assumption is that the regression coefficients corresponding to all non-constant terms in the model can be estimated unbiasedly. A bias is allowed in the estimate for the constant term. Loosely speaking, it means that the estimate of the slope of the regression is unaffected by nonresponse, but the estimate of the level may be different.

In case of simple post-stratification the assumption means that all stratum means have the same bias, i.e. the relative distances between the stratum means are preserved under nonresponse.

Schouten (2003) proves that under this assumption the bias of the generalized regression estimator can be bounded approximately by

$$\frac{S_\rho S_Y}{\overline{\rho}}\left(R_{\rho,Z}R_{Z,Y} - R_{\rho,\tilde{X}\tilde{\beta}_{X,Y}}R_{\tilde{X}\tilde{\beta}_{X,Y},Y} \pm \sqrt{1-R^2_{\rho,Z}}\sqrt{1-R^2_{Z,Y}}\right) \tag{4.3.1}$$

for any variable Z. $\tilde{X}$ is the vector of dummy variables, obtained by removing the dummy corresponding to the constant term. Likewise, $\tilde{\beta}_{XY}$ is the reduced vector of regression coefficients.

Comparing (4.3.1) with (4.2.7) it can be seen that the generalized regression estimator produces a shift of the interval for the bias. The maximal absolute bias of the regression estimator is double the width of (4.3.1) taking $Z = \tilde{X}'\tilde{\beta}_{X,Y}$. Consequently, we must search for a vector of auxiliary variables that minimizes

$$2\frac{S_\rho S_Y}{\overline{\rho}}\sqrt{1-R^2_{\rho,\tilde{X}\tilde{\beta}_{X,Y}}}\sqrt{1-R^2_{\tilde{X}\tilde{\beta}_{X,Y},Y}}. \tag{4.3.2}$$

Since $2\frac{S_\rho S_Y}{\bar{\rho}}$ is independent of the choice auxiliary variables, it suffices to minimize

$$W(X) = \sqrt{1 - R^2_{\rho,\tilde{X}\tilde{\beta}_{X,Y}}}\sqrt{1 - R^2_{\tilde{X}\tilde{\beta}_{X,Y},Y}}. \qquad (4.3.3)$$

This quantity cannot be computed using the available response data. However, Schouten (2003) proves that under the response bias assumption, the correlation $R_{\tilde{X}\tilde{\beta}_{X,Y},Y}$ in the complete sample can be approximated by the correlation $R^*_{\tilde{X}\tilde{\beta}_{X,Y},Y}$ in the response. Therefore, the criterion

$$W^*(X) = \sqrt{1 - R^2_{\rho,\tilde{X}\tilde{\beta}_{X,Y}}}\sqrt{1 - (R^*_{\tilde{X}\tilde{\beta}_{X,Y},Y})^2}. \qquad (4.3.4)$$

is used to determine the weighting model. The proposed selection strategy starts with a simple model with only one weighting variable, namely the variable that minimizes (4.3.4). In subsequent steps iteratively variables are added and removed. Variables are only added or removed in case values of (4.3.4) changes substantially. See Schouten (2003) for further details.

4.4 Example

In this section the proposed selection strategy is applied to the 1998 Integrated Survey on Household Living Conditions (POLS).

From the POLS 1998 survey we selected six important target variables:

- Employment in two classes: employed or unemployed.
- Owner of a house: yes or no.
- At least one activity per month in a club (sports, music, etc.): yes or no.
- Owner of a pc or laptop: yes or no.

Moreover, two auxiliary variables were selected that were also treated as target variables, namely receiving a form of social allowance (yes or no) and ethnic origin (foreign or native). For these two variables, the values in the full sample were available. Therefore, it is possible to see to what extend weighting is capable of completely removing a nonresponse bias.

Table 4.4.1 lists the auxiliary variables used to determine an effective weighting model.

Table 4.4.2 shows the separate steps in the determining the optimal weighting model for the target variable Social allowance. The first variable entering the weighting model is $AgeMar_{36}$, a combination of age and marital status in 36 categories.

Then a second variable is added to the model. It turns out that $HValue_{12}$ (average house value) produces the largest reduction. Since $AgeMar_{36}$ is a crossing of two variables (age and marital status), also the effect of just including one of them has

been investigated. However, replacing $AgeMar_{36}$ by either Age_{15} or $MarStat_4$ does not give a better result. So, $AgeMar_{36}$ is kept in the model.

Table 4.4.1. Available auxiliary variables

Acronym	Description	Values
Sex_2	Sex	male, female
Age_{15}	Age	15 categories
$MarStat_4$	Marital status	4 categories
$AgeMar_{36}$	Combination of age and marital status	36 categories
$EGroup_7$	Ethnic group	7 categories
$Egen_4$	Ethnic generation	4 categories
Job_2	Having a job	yes, no
$SocAllow_2$	Receiving a social allowance	Yes, no
$Province_{12}$	Province of residence	12 provinces
$Region_4$	Region of residence	4 regions
$ProvB_{16}$	Combination of provinces and big cities	16 regions
$Children_2$	Children in the household	Yes, no
$HHType_5$	Type of household	5 categories
$HHSize_5$	Size of the household	5 categories
$Urban_5$	Degree of urbanization	5 levels
$Town_8$	Size of town	8 levels
$District_{27}$	Interviewer district	27 categories
$Phone_2$	Lister phone number	Yes, no
$HValue_{12}$	Average house value	12 categories
$NonNative_8$	Percentage of non-natives	8 categories

Then a third variable is added to the weighting model. It turns out that $Phone_2$ (having a listed phone number) gives the biggest reduction.

Table 4.4.2. Result of selection strategy for Social allowance

Model	$\bar{y}^*_{REG}$	$R^*_{X'\beta^*_{X,Y},Y}$	$R_{\rho,X'\beta^*_{X,Y}}$	$W^*(x)$
$AgeMar_{36}$	0.109	0.34	-0.05	0.940
$AgeMar_{36}$ + $HValue_{12}$	0.112	0.36	-0.08	0.930
$MarStat_4$ + $HValue_{12}$	0.109	0.22	-0.08	0.973
Age_{15} + $HValue_{12}$	0.109	0.33	-0.06	0.943
$AgeMar_{36}$ + $HValue_{12}$ + $Phone_2$	**0.114**	**0.37**	**-0.11**	**0.925**
$HValue_{12}$ + $Phone_2$	0.111	0.16	-0.16	0.975
$MarStat_4$ + $HValue_{12}$ + $Phone_2$	0.111	0.23	-0.11	0.968
Age_{15} + $HValue_{12}$ + $Phone_2$	0.113	0.34	-0.10	0.937

Now that the model contains three variables, it can be investigated whether a variable can be removed without significant loss. This is not possible. Removing $AgeMar_{36}$ causes the criterion value to rise from 0.925 to 0.975. Also replacing $AgeMar_{36}$ by either $MarStat_4$ or Age_{15} leads does not improve the situation. And

adding more auxiliary variables does not result in ignorant improvements. So, the final weighting model for the target variable Social Allowance contains the three variables $AgeMar_{36}$, $HValue_{12}$, and $Phone_2$.

The model selection strategy has been applied to all target variables. The resulting weighting models are listed in table 4.4.3.

Table 4.4.3. Weighting models obtained by selection strategy

Variable	Model
Employment	$Age_{15} + EGroup_7 + Phone_2 + HValue_{12} + Sex_2$
Owner of a house	$HValue_{12} + HHType_5 + NonNative_8 + ProvB_{16} + Age_{15}$
Active in a club	$Age_{15} + NonNative_8 + Phone_2$
Owner of a pc	$Age_{15} + HValue_{12} + EGroup_7 + HHSize_5 + SocAllow_2$
Educational level	$Age_{15} + NonNative_8 + ProvB_{16} + HValue_{12}$
Religion	$EGroup_7 + ProvB_{16} + Phone_2 + Job_2$
Social allowance	$AgeMar_{36} + HValue_{12} + Phone_2$
Ethnic background	$NonNative_8 + Phone_2$

A number of conclusions can be drawn from these results:

- Different target variables need different weighting models. So, effective correction for non-response requires several set of weights to be included in the survey data file;
- Traditional demographic variables are not sufficient for weighting. More auxiliary variables are needed. Two important variables are $Phone_2$ and $HValue_{12}$. They appear in five of the eight weighting models. A substantial role is also played by variables related to ethnicity ($NonNative_8$ and $EGroup_7$);
- The models in table 4.4.3 contain no interaction terms. Crossing auxiliary variables does not lead to significant improvements in estimates. This is a fortunate conclusion. Using only marginal distributions reduces the risk of empty cells.

Figure 4.4.4. Resulting estimates for various weighting models

Variable	Response mean	Current model	New model	Combined model	Full sample mean
Employment	57.7 %	58.2 %	57.4 %	57.4 %	
Owner of a house	63.3 %	60.4 %	59.4 %	59.0 %	
Active in a club	46.6 %	45.6 %	45.1 %	44.8 %	
Owner of a pc	59.8 %	58.3 %	57.2 %	57.3 %	
Social allowance	10.4 %	11.0 %	11.4 %	11.5 %	12.1 %
Ethnic background	12.5 %	13.3 %	14.4 %	15.0%	15.0 %

Table 4.4.4 gives some indication how effective the various weighting models are. The current model denotes the weighting model currently used in POLS:

$$(Sex_2 \times Age_3 \times MarStat_2) + (Sex_2 \times Age_{15}) + (Region_4 \times Age_3) + Urban_5 + Region_{16} + HHSize_5 + MarStat_4$$

It only contains the traditional demographic variables. The new model denotes the model obtained with the selection strategy, see table 4.4.3.

For the two variables social allowance and ethnic origin also the estimates based on the full sample are available. By looking at the bottom two rows of the table, it becomes clear that the new weighting performs better than the current model. However, it also becomes clear that these models are not completely capable of removing the bias.

For nearly all target variables (with the exception of employment) a trend can be observed in the estimates. Using the current weighting model causes the estimate to shift in a certain direction. The new model causes a more profound shift in the same direction. This is an indication that the new model performs better than the current model.

As was mentioned earlier, different target variables require different weighting models. This is inconvenient, because it leads to several sets of weights to be included in the survey data file. One solution of this problem could be to attempt to combine all weighting models in table 4.4.3 to one general model. One possible model might be:

$$Age_{15} + HValue_{12} + Phone_{12} + NonNatives_8 + EGroup_7 + ProvB_{16} + HHType_5$$

Two auxiliary variables were not included: social allowance and sex. These variables gave only small changes and were selected only once. Furthermore, household type was included. This variable only appeared in one model. Due to the strong relationship it can function as a substitute for household size. Finally, the variable Age_{15} was included and not $AgeMar_{36}$ because marital status only once had a significant impact. Table 4.4.4 also contains the estimates based in this combined model. The estimates for employment, ownership of a PC, and social allowance (almost) did not change. The estimates for active in a club and house ownership still show some improvement. One may wonder why this model was not found by the selection strategy. The reason is that the improvement is the sum of a number of non-significant contributions, and moreover, the strategy attempts to keep the model as small as possible.

Note that the estimate for ethnic background is identical to the full sample mean. This comes as no surprise as this variable is included as an auxiliary variable in the weighting model.

The combined weighting model shares the variables age and region with the current weighting model. New are average house value, percentage of non-natives, ethnic group, household type and having a listed telephone number. By comparing

estimates for the two models, one can see that the largest difference is found for percentages of house owners (1.4%) and PC owners (1.0%).

When we compare the unweighted response means with estimates based on the combined model, we see that the largest shift is 4.3%, namely for the percentage of house owners. All other shifts are smaller, but still substantial.

4.5 Conclusion

Biases that were encountered in the POLS 1998 data, could be reduced by the use of weighting models. However, they did not disappear completely. This is somewhat disappointing, because it means that the available auxiliary variables are not capable of completely explaining response behaviour.

In all investigated situations we found that weighting models with interaction effects give estimates very similar to weighting models with only main effects. This means we can substantially reduce the number of parameters in the model without affecting the outcomes and hence incorporate more auxiliary information.

For the two auxiliary variables that we regarded, i.e. having a form of social allowance and ethnic background, selected weighting models produce estimates that are much closer to the sample means than the response means.

The shifts produced by weighting models relative to the response means are usually not very large. The largest shift that we found is 4.3% for ownership of a house. The percentage of house owners dropped from 63.3% to 59.0%.

The differences between the combined weighting model and the current weighting model for POLS may run up to more than 1%. This seems especially due to the new auxiliary variables used: average house value, percentage of non-natives, and having a listed telephone number. These three variables seem to be very interesting for use in weighting models.

Analysis of the nonresponse of POLS 1998 revealed that the auxiliary variables interviewer district and degree of urbanization relate quite strongly to response behaviour. However, these variables did not turn out to be very interesting for weighting nonresponse. It seems that these variables hardly relate to the important survey questions.

Our finding seems to indicate the usefulness of the proposed strategy for selecting auxiliary variables in weighting models. This strategy focuses simultaneously on the relation between auxiliary variables and survey questions, and auxiliary variables and response behaviour. Auxiliary variables are only interesting in case both relations exist.

Another benefit of the strategy is that the construction of strata can be done in one step. In this report we used post-stratification as a method to adjust for bias. However, the selection strategy may equally well be used to form cells in propensity score weighting.

There are still a number of issues that need to be resolved. First, we need to investigate to what extent correlations between target and auxiliary variables are affected by nonresponse.

Second, in practice it is not very convenient to have a weighting model for each survey question. For this reason the construction of weighting models is often solely based on the prediction of the response behaviour. It is then believed that there will always be at least one survey question that relates to the auxiliary variables that best explain the nonresponse mechanism. As we indicated before this approach may lead to weighting models containing excess variables. Alternatively, one may select the most important survey questions or do a principal component analysis and take the first, say five, components. See for instance Geuzinge, van Rooijen and Bakker (2000). Still it remains unclear how to construct a weighting model with the combined effect of several weighting models.

Another important aspect is the assumption underlying the response mechanism. The selection strategy was based on the assumption that stratum response means are biased but in the same direction. As a consequence relative distances between stratum means are preserved under nonresponse. There seems to be some evidence that this assumption is closer to the truth than the MAR assumption. More evidence is needed, however.

In this paper we did not pay attention to the variance of estimators. We believe it to be less important than bias. Particularly for large sample sizes, the bias will be much larger than the standard error of the estimators. However, variance does play a role in the quality of estimators. However, computation of variances in the nonresponse assumption is not straightforward.

References

Bethlehem, J.G. (1988), Reduction of the nonresponse bias through regression estimation. *Journal of Official Statistics* 4, pp. 251-260.

Bethlehem, J.G. (1996), *Bascula for Weighting Sample Survey Data, Reference Manual*. Statistics Netherlands, Voorburg, The Netherlands.

Bethlehem, J.G. (2002), Weighting nonresponse adjustments based on auxiliary information. In: R.M. Groves, D.A. Dillman, J.L. Eltinge, and R.J.A. Little (eds.), *Survey Nonresponse*. Wiley, New York.

Bethlehem, J.G. and Kersten, H.M.P. (1985), On the treatment of nonresponse in sample surveys. *Journal of Official Statistics* 1, pp. 287-300.

Bethlehem, J.G. and Schouten, B. (2004), *Nonresponse Analysis of the Integrated Survey on Living Conditions (POLS)*. Discussion Paper 0230. Statistics Netherlands, Voorburg, The Netherlands

Czajka, J.L., Hirabayashi, S.M., Little, R.J.A., and Rubin, D.B. (1992), Projecting from advance data using propensity modelling: An application to income and tax statistics, *Journal of Business & Economic Statistics*, 10, 117–131.

Deville, J.C. and Särndal, C.E. (1992), Calibration estimators in survey sampling. *Journal of the American Statistical Association* 87, pp. 376-382.

Ekholm, A. and Laaksonen, S. (1991), Weighting via response modelling in the Finnish household budget survey, *Journal of Official Statistics*, 7, 325–337.

Eltinge, J.L. and Yansaneh, I.S. (1997), Diagnostics for formation of nonresponse adjustment Cells, with an application to income nonresponse in the U.S. Consumer Expenditure Survey, *Survey Methodology*, 23, 33–40.

Everaers, P. & Van der Laan, P. (2001), *The Dutch virtual Census*. E-Proceedings of the 53th Session of the International Statistical Institute, Seoul, Korea.

Gelman, A. and Carlin, J.B. (2000), Poststratification and weighting adjustments. In: Groves, R.M., Dillman, D.A., Eltinge, J.L., Little, R.J.A. (Eds), *Survey Nonresponse*. Wiley, New York.

Geuzinge, L., Rooijen, J. van, and Bakker, B.F.M. (2000), The use of administrative registers to reduce non-response bias in household surveys, *Netherlands Official Statistics* 2000-2, 32–39.

Hansen, M.H. and Hurvitz, W.N. (1946), The problem of nonresponse in sample surveys. *Journal of the American Statistical Association* 41, pp. 517-529.

Holt, D. and Smith, T.M.F. (1979), Post Stratification. *Journal of the Royal Statistical Society*, A, 142, pp. 33-46.

Isaki, C.T. and Fuller, W.A. (1982), Survey design under the regression superpopulation model. *Journal of the American Statistical Association* 77, pp. 89-96.

Kalton, G. and Maligalig, D.S. (1991), A comparison of methods of weighting Adjustment for Nonresponse. *Proceedings of the 1991 Annual Research Conference.* U.S. Bureau of the Census, pp. 409-428.

Kersten, H.M.P. and Bethlehem, J.G. (1984), Exploring and reducing the nonresponse bias by asking the basic question. *The Statistical Journal of the United Nations Commission for Europe* 2, pp. 369-380.

Knottnerus, P. (2003), *Sample Survey Theory, Some Pythagorean Perspectives*. Springer-Verlag, New York.

Little, R.J.A. (1986), Survey nonresponse adjustments for estimates of means, *International Statistical Review* 54, pp. 139–157.

Little, R.J.A. (1993), Post-stratification: A modeler's perspective. *Journal of the American Statistical Association* 88, pp. 1001-1012.

Mesa, D.M., Tsai, P. and Chambers, R.L. (2000), *Using Tree-Based Models for Missing Data Imputation: An Evaluation Using UK Census Data*. University of Southampton.

Robinson, P.M. and Särndal, C.E. (1983), Asymptotic properties of the generalized regression estimator in probability sampling. *Sankhya* B 45, pp. 240-248.

Rosenbaum, P.R. and Rubin, D.B. (1984), Reducing bias in observational studies using subclassification on the propensity score, *Journal of the American Statistical Association*, 79 (387), pp. 516–524.

Schouten, B. (2003), *Reduction of Nonresponse Bias using Auxiliary Variables*. Report 1814-02-TMO. Statistics Netherlands, Methods and Informatics Department, Voorburg, The Netherlands.

Schouten, B. and Bethlehem, J.G. (2002), *Respondents, persuadable non-respondents and persistent nonrespondents*. Research Paper 0230. Statistics Netherlands, Methods and Informatics Department, Voorburg, The Netherlands.

Thomsen, I. (1973), A note on the efficiency of weighting subclass means to reduce the effects of non-response when analyzing survey data. *Statistisk Tidskrift* 11, pp. 278-283.

Voogt, R. (2004), *I'm not interested – Nonresponse Bias and Stimulus Effects in Election Research*. PhD. Thesis, University of Amsterdam.